Explainability and case-based reasoning

Olena Verbytska

June 28, 2024

The General Data Protection Regulation laws in the EU state that all people have the right to "meaningful information about the logic behind automated decisions using their data" [7]. This means that organizations using automated decision-making processes must provide clear and understandable explanations of how these decisions are made, ensuring transparency and accountability in the use of personal data. For example, suppose an AI system predicts whether a patient has a high risk of developing a certain disease. Explainability would involve not just giving a risk score, but also detailing which factors (e.g., age, medical history, genetic markers) influenced this prediction and how each factor contributed to the final decision.

1 Explainable vs Interpretable AI

For Explainable AI systems it's hard to know **how** a result was arrived at, but you mostly know **why**, if you trust the explanation.

Strictly Interpretable AI lacks explanations. It's very easy to see **how** the algorithm arrived at its conclusion, but not **why** the criteria it used are sensible.

Explainability and interpretability are often seen as binary categories. However, this perspective is misleading as they should be viewed as a spectrum, reflecting varying degrees and types of understanding provided by different models and methods [4].

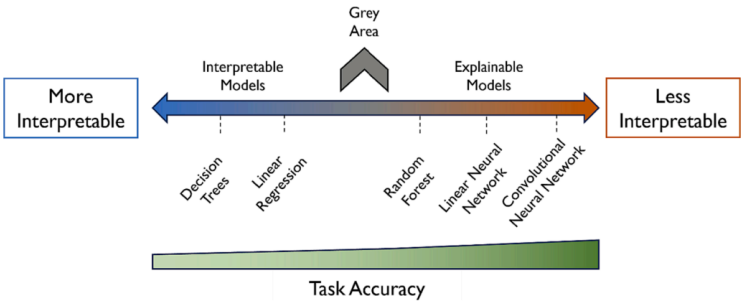

Figure 1: Interpretability and accuracy trade-off [4]

2 Approaches to achieve explainability

The scientific community has focused on deep learning (DL) due to its versatility, high performance, high generalization capacity, and multidisciplinary applications, among many other qualities. The DL approach enhances the performance of previous works based on machine learning (ML) models, particularly in biomedical image classification and segmentation. However, both systems are often viewed as black boxes, providing predictions without explaining the underlying decision-making process. Consequently, users such as healthcare practitioners struggle to understand and verify the results produced by ML or DL models. This lack of transparency has led to relatively low acceptance of AI in the healthcare field, as practitioners still rely heavily on evidence-based diagnoses to guide treatment decisions.

Therefore, an explanation and interpretation of the model's output and work are required to improve the use of these systems in various clinical applications. Specifically, Explainable Artificial Intelligence (XAI) is defined as the set of features that explain how the AI model constructed its prediction. In this section we will show a short review of the most popular XAI based on a Systematic Review of 2011-2022 years [6] and original papers.

There exist five XAI approaches: visual, numerical, rule-based, textual, and example-based.

2.1 Visual approaches

2.1.1 GradCAM (Gradient-weighted Class Activation Mapping)

Description: Provides visual explanations by emphasizing the significant regions in an image that impact the predictions of a convolutional neural network (CNN), either without fully-connected layers or prior to their application.

Application: Primarily used in computer vision tasks rather than in the conventional ML models. GradCAM's final output is a heatmap, highlighting the regions the model considers to be important for the final prediction.

2.1.2 LRP (Layer-wise Relevance Propagation)

Description: While GradCAM is applied directly to the last convolutional layer of a CNN model, LRP works backward by first calculating the relevance score for a specific output at the model's classifier layer, which is also the last layer. It decomposes the prediction into contributions of individual input features by redistributing the output score layer by layer until the input layer is reached.

Application: Commonly used in image classification tasks to produce heatmaps that highlight the parts of the input (e.g., pixels in an image) that are most relevant for a specific prediction, providing a visual representation of the model's decision process. However, LRP appears to be less popular than GradCAM for heatmap generation due to the complexity of implementing the algorithm.

2.2 Numerical approaches

2.2.1 SHAP (SHapley Additive exPlanations)

Description: Uses Shapley value from cooperative game theory to assign each feature an importance value for a particular prediction.

Application: Applicable across various domains to quantify the contribution of each feature in a model's prediction.

Consider a group of players working together to clear a game. How would the final reward be split if each player contributed differently? Shapley value can be used in this situation to guarantee that the allocation of rewards to each player is fair by calculating the marginal contribution of each player. Shapley values satisfy the four axioms for calculating each player's marginal contribution, which is listed below.

1. Efficiency: the final reward must be shared among the players in cooperation.

2. Symmetry: players who made the same contribution as each other will receive the same amount of reward.

3. Dummy: players who did not contribute to the game clearance are known as dummy players and will receive no reward.

4. Additivity: if the game has multiple parts, the player's reward allocation must take into account the individual contribution to each part rather than the collective contribution to the game as a whole.

As a result, SHAP has the potential to be used in healthcare by analyzing the contribution of biomarkers or clinical features (players) to a specific disease outcome (reward).

2.2.2 Feature Importance

Description: Provides a ranking of features based on their impact on the model's predictions.

Application: Common in tree-based models like Random Forests and Gradient Boosting Machines to identify which features are most influential.

2.3 Rule-based approaches

2.3.1 Rule-based Systems

Description: Uses a set of predefined "if-then" rules to make decisions or inferences. Global explanations attempt to make the entire model transparent, whereas a local explanation does not aim to explain the entire model; instead, it changes the input features and observes how these affect the output prediction.

Application: Widely used in expert systems and decision support systems where transparent and understandable logic is crucial. For example, they are employed in medical diagnosis systems to assist doctors in identifying diseases based on symptoms.

2.3.2 EBM (Explainable Boosting Machine)

Description: Combines generalized additive models (GAMs) with boosting techniques to create models that are both accurate and interpretable through understandable rules. EBM is an ML model on its own rather than a function that was added to an ML or DL model. Results in some number of small tree models for each feature.

Application: Suitable for high-stakes domains like healthcare and finance where clear decision rules are necessary.

2.3.3 Fuzzy classifiers

Description: A classifier based on fuzzy logic that deals with reasoning that is approximate rather than fixed and exact. It uses fuzzy sets and rules to model uncertainty and imprecision in data. Fuzzy logic mimics human decision-making by taking into account the various ranges of possibilities between "yes" and "no," such as "definitely yes" or "maybe no".

Application: Used in areas requiring human-like reasoning, such as pattern recognition, control systems, and decision-making processes where data may be uncertain or imprecise. The truth value given by fuzzy logic should not be confused with the probability score generated by a SoftMax classifier which measures the likelihood of an event occurring.

2.4 Textual approaches

2.4.1 LIME (Local Interpretable Model-agnostic Explanations)

Description: Provides local explanations by approximating the model locally with an interpretable model and expressing the reasoning in human-readable text. LIME can only explain a single sample in the dataset, therefore, sometimes SHAP is preferred over LIME to explain the entire data set. Also, SHAP provides the results that are easier to interpret. LIME aims to understand decisions at the individual level by altering the input example and determining which changes are most likely to affect the decision. For image analysis, this involves occluding sections of the image, with the explanation provided as a heatmap indicating the components of the image that were most important in the decision-making process.

Application: Used for explaining individual predictions of complex models in natural language.

2.5 Example-based Approaches

2.5.1 CBR (Case-Based Reasoning)

Description: Solves new problems by adapting solutions that were used to solve past, similar problems. The CBR's weakness is its reduced performance when too many cases are stored in its system.

Application: Utilized in diagnostic systems and customer support to provide explanations based on previous similar cases.

2.5.2 Prototype and Criticism Models

Description: Uses representative examples (prototypes) and contrasting cases (criticisms) to explain predictions. These prototypes exemplify the essential characteristics that define a class or a decision. In contrast, criticism examples are instances that deviate significantly from the prototype or are misclassified. By examining both prototypes and criticisms, this approach aims to illuminate why a model makes specific predictions or decisions.

Application: Applied in machine learning tasks to help users understand the decision-making process by comparing with known examples.

3 Problems

1. People tend to over-trust AI especially in medical systems [1].
2. The human tend is to ascribe a positive interpretation. It is so-called confirmation bias.
3. The interpretability gap of explainability methods relies on humans to decide what a given explanation might mean. Often the responsible person (e.g. doctor, nurse, medical staff) do not have enough competence in AI to explain model behavior correctly.
4. Sometimes even the hottest parts of the heatmap contain both useful and non-useful information, and simply localising the region does not reveal exactly what it was in that area that the model considered useful [7].
5. Untrained networks can produce heatmaps that look reassuring [5].
6. Common visual explanations (such as GradCAM or LRP) remain unchanged even when precise modifications are made to the input that change the model's predictions [2].
7. Explanations have no performance guarantees.
8. Explainability tools do not allow fully control prejudgemental behaviour of AI, e.g. discriminatory policies against women and minority ethnic groups [3].

4 Solutions

- While looking at heatmaps the users should question themselves **not where** the model was looking, but instead **whether it was reasonable** that the model was looking in this region.
- Explainability techniques do not work as a human may expect for producing valid, local explanations to justify the use of model predictions. It is more realistic to view these methods as global descriptions of how a model functions.
- The only effective way to justify the decisions of AI systems is thorough, careful, meticulous safety and validation efforts by using XAI.
- Enhance the quality of test sets used to evaluate clinical diagnostic models to ensure they accurately represent real-world scenarios, reducing distractions from irrelevant image regions.
- Advocate for rigorous validation across diverse and distinct populations to demonstrate improved patient outcomes and mitigate disproportionate impacts on marginalized groups.
- The users or subjects of AI are not the best interpretators of explainability techniques.
- Use explainability techniques for analysis and as supplements to algorithmic evaluations, directing explanations towards developers, auditors, and regulators to improve transparency and accountability in AI systems.
- Randomised controlled trials have historically been the gold-standard way to evaluate medical interventions, and it should be no different for AI systems, as it can help to improve and trust them without even knowing how they work.

References

[1] Vinay Chamola et al. "A Review of Trustworthy and Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 11 (2023), pp. 78994–79015. DOI: 10.1109/ACCESS.2023.3294569.

[2] Jindong Gu and Volker Tresp. "Saliency Methods for Explaining Adversarial Attacks". In: *arXiv* (2019). DOI: https://doi.org/10.48550/arXiv.1908.08413.

[3] Zhang H et al. "Hurtful words: quantifying biases in clinical contextual word embeddings". In: *Proceedings of the ACM conference on health, inference, and learning* (2020).

[4] Abrantes J. and Rouzrokh P. "Explaining explainability: The role of XAI in medical imaging". In: *European Journal of Radiology* (2024). DOI: https://doi.org/10.1016/j.ejrad.2024.111389.

[5] Adebayo J. et al. "Sanity checks for saliency maps". In: *Adv Neural Inf Process Syst* 31 (2018).

[6] Hui Wen Loh et al. "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)". In: *Computer Methods and Programs in Biomedicine* 226 (2022), p. 107161. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2022.107161.

[7] Ghassemi M., Oakden-Rayner L., and Beam A. L. "The false hope of current approaches to explainable artificial intelligence in health care". In: *The Lancet Digital Health* 3 (Nov. 2021), e745–e750. DOI: 10.1016/S2589-7500(21)00208-9.